Published in final edited form as:
Proc Int Conf Spok Lang Process. 2002 ; 2002: 1689–1692.

AUDIOVISUAL INTEGRATION OF SPEECH BY CHILDREN AND ADULTS WITH COCHEAR IMPLANTS

Karen Iler Kirk[1], **David B. Pisoni**[1,2], and **Lorin Lachs**[2]

Karen Iler Kirk: kkirk@iupui.edu

[1]Department of Otolaryngology-HNS, Indiana University School of Medicine, Indianapolis

[2]Department of Psychology, Indiana University, Bloomington

Abstract

The present study examined how prelingually deafened children and postlingually deafened adults with cochlear implants (CIs) combine visual speech information with auditory cues. Performance was assessed under auditory-alone (A), visual- alone (V), and combined audiovisual (AV) presentation formats. A measure of visual enhancement, R_A, was used to assess the gain in performance provided in the AV condition relative to the maximum possible performance in the auditory-alone format. Word recogniton was highest for AV presentation followed by A and V, respectively. Children who received more visual enhancement also produced more intelligible speech. Adults with CIs made better use of visual information in more difficult listening conditions (e.g., when mutiple talkers or phonemically similar words were used). The findings are discussed in terms of the complementary nature of auditory and visual sources of information that specify the same underlying gestures and articulatory events in speech.

1. INTRODUCTION

Cochlear implants are electronic auditory prostheses for individuals with severe to profound hearing impairment that enable many of them to perceive and understand spoken language. However, the benefit to an individual user varies greatly. Some CI users can communicate successfully over a telephone even when lipreading cues are unavailable whereas others find that the CI helps them understand speech only when visual information also is available. One source of variability may result from the way in which these initial sensory inputs are coded and processed by higher centers in the auditory system. For example, listeners with detailed knowledge of the underlying phonotactic rules of English may be able to use limited or degraded sources of sensory information in conjunction with this knowledge to achieve better overall performance. Fortunately, daily speech communication is not limited to input from only one sensory modality. Optical information about speech obtained from lipreading improves speech understanding in listeners with normal hearing [1] as well as persons with CIs [2]. In the following two experiments, we examined the ability of profoundly deaf individuals to integrate the auditory information from a CI with visual speech cues.

2. EXPERIMENT I

This experiment examined audiovisual enhancement in children with cochlear implants and it's relationship to spoken language processing and speech intelligibility.

2.1. Participants

Twenty-seven children with prelingual deafness (onset before 3 years) who had used a CI for at least two years participated. Their average age at onset of deafness was 0.51 years and their average age at implantation was 4.52 years. All of the children used a Nucleus CI. Fifteen children used Total Communication (TC) (combined signed and spoken English). The remaining 12 children used oral/aural communication (OC).

2.2 Methods

Children were administered a sentence test, The Common Phrases Test under three presentation conditions, A, V and AV. The test was administered live voice. During A only presentation, the experimenter's face was obscured by a cloth mesh screen. During V only presentation, the child's CI was removed or turned off. Ten different phrases were presented in each condition. Performance in each condition was scored by the percent of phrases correctly repeated, in their entirety by the child.

In addition to the above measures, children were administered two 50-item tests of monosyllabic word recognition using the Lexical Neighborhood Test (LNT) and the Phonetically Balanced Kindergarten word lists, and a 24-item list of two-three syllable words, the Multisyllabic Lexical Neighborhood Test (MLNT). Both the LNT and MLNT contain lexically-controlled word lists. That is, half of the tokens on each test are lexically easy, in that they occur often in the language and have few phonemically similar words with which they can be confused. The remaining items on the LNT and MLNT are lexically hard, in that they occur less often in the language and have many similar words with which they can be confused. Performance on these measures was scored as the percent of words correctly identified. These measures of spoken word recognition were chosen because they are among the most commonly used to determine CI candidacy and to monitor postimplant outcomes in children.

A measure of receptive language also was obtained to assess differences in the ability of these children to use language in general. The Peabody Picture Vocabulary Test (PPVT) is a standardized test that provides a measure of receptive language development based on word knowledge. These test items were administered using the child's preferred communication mode, either TC or OC.

In addition to these receptive measures of performance, a test of speech production was administered to each child to obtain a measure of their speech intelligibility. Each child imitated 10 sentences; their utterances were recorded and played back later for transcription by three naive adult listeners who were unfamiliar with the speech of deaf talkers. Speech intelligibility scores were measured by calculating the average number of words correctly identified by the panel of listeners.

2.3 Results and Discussion

Each child's score on the Common Phrases Test in the A and AV conditions were combined to obtain the measure R_A, which indexes the relative gain in speech perception due to the addition of visual information about articulation [1], R_A was computed using the following formula

$$R_A = (AV - A)/(100 - A) \quad (1)$$

where AV and A represent the accuracy scores obtained in the audiovisual and auditory-alone conditions, respectively. From this formula, one can see that R_A measures the gain in accuracy in the A condition, normalized relative to the amount by which speech recognition scores could possibly improve above auditory-alone scores.

Table 1 presents the average performance of the children in the three presentation formats of the Common Phrases Test along with the average R_A. The range of scores under all presentation formats varied considerably. In the A condition, scores varied from 0%–90% correct. Similarly, in the V condition, scores ranged from 0% to 80% correct. Scores in the AV condition varied across the entire possible range. It is important to note that there was no significant difference between scores in the two unimodal conditions, A and V. Thus, there was no overall tendency for these children to rely more on one input modality than another. Inspection of the R_A scores revealed that children with CIs exhibited a wide range in their ability to combine multisensory inputs. For 23 of the children, their AV score was significantly higher than the score they obtained in the A condition.

There also were significant correlations among the three presentation conditions, A was correlated with V ($p < .05$); AV was correlated with A ($p<.01$) and AV was correlated with V ($p<.01$). These correlations suggest a common underlying source of variance; the same set of skills may be used on the Common Phrases Test regardless of presentation modality. However, these relations may not be due simply to a more global language proficiency or to an ability to use the contextual framework of the Common Phrases Test. Correlations between Common Phrases scores in all three presentation formats and the PPVT age equivalent scores were not significant.

We also analyzed the relationship between R_A and PPVT age equivalence scores. Despite the fact that vocabulary knowledge was not related to Common Phrases scores in each presentation format, there was a relationship between R_A and PPVT age equivalence ($p <. 05$). This indicates that the ability to benefit from combined audiovisual input is related to global language abilities, independent of the child's absolute perception scores.

Performance on the Common Phrases Test in the auditory-alone presentation format was significantly related to performance on all three spoken word recognition measures ($p<.05$). Correlations also were computed between audiovisual enhancement scores, R_A and the auditory-only spoken word recognition measures, and between R_A and the speech intelligibility scores (See Table 2). Audiovisual enhancement was significantly correlated with spoken word recognition on the MLNT and the LNT Easy words; lack of significant correlations with the LNT Hard words and PBK words was likely due to floor effects on

those difficult measures. Audiovisual enhancement also was significantly correlated with the children's speech production skills as measured by the speech intelligibility task.

The results revealed that the children's skills in deriving benefit from audiovisual sensory input are not independent but are closely related to auditory-alone spoken word recognition and speech production, both of which draw on a common set of underlying phonological processing abilities. These skills include perceptual, cognitive, and linguistic processes that are used in the initial encoding, maintenance, rehearsal and manipulation of the phonological and lexical representations of spoken words, and the construction and implementation of sensory-motor programs for speech production and articulation. The links between the receptive and expressive aspects of language reflect the child's developing linguistic knowledge and use of phonology, morphology and syntax and his or her attempts to use this knowledge productively in a range of language processing tasks.

3. EXPERIMENT II

In daily activities, listeners with CIs perceive speech under a wide variety of conditions, including face-to-face conversation, televsion, and over the telephone. Success in recognizing words and understanding speech may differ substantially under such diverse listening conditions. This study examined the ability of postlingually deafened adults to integrate the limited auditory information they receive from a CI with visual speech cues when stimulus variability in the form of different talkers or lexical characteristics is present.

3.1 Participants

Forty-one adults served as listeners in this study and were paid for their participation. Twenty were postlingually deafened adult users of CIs who were recruited from the clinical population at Indiana University. All listeners with CIs had a profound bilateral sensorineural hearing loss and had used their CI for at least six months. Their mean age at time of testing was 50 years. The control group consisted of 21 adult listeners who were recruited from within Indiana University and the associated campuses; their average age was 42 years. All of the listeners in the comparison group had pure tone thresholds below 25 dB HL at 250, 500, 1000, 2000, 3000, and 4000 Hz and below 30 dB HL at 6000 Hz. Each participant was reimbursed for travel to and from testing sessions and was paid $10.00 per hour of testing.

3.2 Methods

Stimulus materials were drawn from a database of digitally recorded audiovisual speech tokens containing 300 monosyllabic English words produced by five male and five female talkers. For the present study, we created six equivalent word lists that would allow us to examine the effect of presentation format, talker variability, and lexical competition on spoken word recognition. Each test list contained 36 words. On each list, half of the words were lexically easy, and half were lexically hard. Two versions of each of the six original word lists were produced: one version contained tokens produced by a single talker. The second version contained tokens produced by six different talkers. This arrangement enabled us to administer a single-talker or multiple-talker version of each test list.

Testing was conducted in a single-walled sound treated IAC booth (Model #102249). The digitized audiovisual stimuli were presented to participants using a PowerWave 604 (Macintosh compatible) computer equipped with a Targa 2000 video board. All listeners were tested individually, one at a time. The experimental procedures were self-paced. Video signals were presented with a JVC 13U color monitor. Speech tokens were presented via a loudspeaker at 70 dB SPL (C weighted) for participants using CIs. Each participant was administered three single talker and three multiple talker lists. Within each talker condition, one list was presented using an auditory-alone format, one using a visual-alone format, and one using an auditory plus visual format. Visual-alone conditions were achieved by attenuating the loudspeaker and auditory-alone conditions were achieved by turning off the video display monitor.

Normal hearing participants were tested using a −5 dB signal to noise ratio in speech spectrum noise at 70dB SPL relative to the 65 dB SPL speech tokens. This SNR was chosen during preliminary testing to prevent most of the participants with normal hearing from attaining ceiling performance on the task. All of the participants were asked to verbally repeat the word that was presented aloud. The experimenter subsequently recorded their responses into computer files online. No feedback was provided.

3.3 Results and Discussion

Table 3 presents a summary of the raw scores obtained by the two groups as a function of presentation format, lexical difficulty, and talker variability. A significant main effect of Presentation Mode was observed for both groups. Regardless of group membership, performance in the visual-alone condition was worse than in the auditory-alone condition, which was even worse than in the audiovisual condition. Because CI and control participants were tested under identical conditions only when visual-only stimuli were presented, direct comparison of performance between the two groups is valid only for this test condition. CI users obtained higher scores in the visual-alone condition than their normal-hearing counterparts. This is not surprising given that adults with hearing impairment have experience in utilizing visual speech cues to supplement the information they receive through the auditory channel. These findings are consistent with a recent report by Bernstein, Auer, and Tucker [3] who found reliable differences in the performance of normal hearing and hearing impaired speechreaders on a visual-alone speech perception task.

Speech intelligibility scores obtained under each presentation format were correlated separately for each group of listeners. Significant correlations were observed between auditory-alone performance and audiovisual performance for both groups of listeners, $r(20) = +0.81$, $p<.001$, for CI listeners and $r(21) = +0.67$, $p<0.001$, for NH listeners. However, the correlations between visual-alone performance and audiovisual performance were not significant for either group. Additional correlations were computed between the auditory-alone and visual-alone performance for each group of listeners. None of these correlations was significant.

The main effect Talker Variability was significant for both the CI and control groups. Overall, single talker lists were identified better than multiple talker lists. Talker Variability

also interacted with Presentation Mode for both groups, although the effect was marginally significant for the CI group. The results on the effects of talker variability are consistent with the proposal that repeated exposure to a single talker allows the listener to encode voice-specific attributes of the speech signal. Once internalized, voice-specific information can improve word recognition performance [4], The “single talker advantage” appears to be most helpful when there is a great deal of lexical competition among words and fine phonetic discrimination is required, as with lexically hard words. Talker-specific information appears to be used in conditions where a detailed perceptual representation of the acoustic/phonetic input can serve to more clearly disambiguate multiple word candidates from within the lexicon. For both groups of listeners, this detail is provided in the audiovisual condition.

For both groups of listeners, lexically easy words were recognized better than lexically hard words, indicating that normal hearing and CI listeners organize and access words from lexical memory in fundamentally similar ways. Thus, phonetically similar words in the mental lexicons of CI users compete for selection during word recognition. This process also is affected by word frequency, such that more frequently occurring words are more apt to win out among phonetically similar competitors. The finding that lexical competition affected the CI group is not surprising because the participants in this group were all post-lingually deafened and had no evidence of any central nervous system involvement prior to or after the onset of deafness. Presumably, they developed robust lexical representations when they had normal hearing and retained some form of this information over time after their hearing loss.

To assess visual enhancement, R_a was calculated for all 41 participants based on the recognition scores obtained in the audiovisual and auditory-alone conditions using Equation 1. R_a was calculated separately for lexically easy and lexically hard words in each of the two talker conditions (see Table 4). Because R_a normalizes for auditory-alone performance, it is possible to compare across listener groups. Overall, R_a was larger for single talker than for multiple talker conditions. The interaction between Talker and Group also was significant. This interaction was due to a difference in visual enhancement for single vs. multiple talker lists that was significant for CI users, $p \leq 0.006$, but not for normal-hearing participants.

Visual enhancement scores for lexically easy words were significantly higher than for lexically hard words. This result indicates that listeners obtained somewhat greater visual benefit from words that have less competition than from words that have more competition. No other main effects or interactions from the R_a ANOVA were significant.

There was no effect of talker variability on visual enhancement for normal-hearing listeners. This finding does *not* mean that NH listeners were unaffected by talker variability. However, talker variability did not affect the degree to which normal-hearing listeners could *combine* audiovisual information. The present findings suggest that CI users are better able to extract idiosyncratic talker information from audiovisual displays than NH listeners are, perhaps because they rely more on visual speech information to perceive speech in every day situations. With repeated exposure to audiovisual stimuli spoken by the same talker, the CI users exhibited a gain above and beyond that observed in normal hearing listeners. Because

NH listeners can successfully process spoken language by relying entirely on auditory cues, they may not have learned to utilize visual cues as successfully [3], For NH listeners, combined audiovisual information from a single talker may not provide any additional information about that talker than the cues provided by auditory-alone presentation.

4. CONCLUSIONS

The present study demonstrates that both prelingually deafened children and postlingually deafened adults can use the degraded auditory input they receive from a CI to supplement visual speech cues they receive from the talker's face. Furthermore, adults with CIs appear to make better use of visual information in more difficult listening conditions when there is ambiguity about the talker or when they are required to make fine phonetic discriminations among acoustically confusable words. The deaf listeners combined auditory and visual speech cues to support open-set word recognition but they do this in somewhat different ways than normal hearing listeners. Intervention and treatment programs that are designed to increase receptive and/or production skills in hearing impaired listeners may wish to emphasize the inherent cross-correlations that exist between auditory and visual sources of speech information.

Acknowledgments

This work was supported by NIH-NIDCD grants K23 DC00126, R01 DC00064, T32 DC00012, and R01 DC00423 and by Psi Iota Xi national sorority.

References

1. Sumby WH, Pollack I. Visual contribution of intelligible speech in noise. JASA. 1954; 26:212–215.
2. Tyler RS, Opie JM, Fryauf-Bertschy H, Gantz BJ. Future directions for cochlear implants. JSLHR. 1992; 16:151–163.
3. Bernstein LE, Auer J, Tucker PE. Enhanced speechreading in deaf adults: Can short-term training/practice close the gap for hearing adults. JSLHR. 2001; 44:5–18. [PubMed: 11218108]
4. Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Percept & Psychophys. 1998; 60:355–376.

Table 1

Mean Common Phrases and R_A score (bounds are associated with the 95% confidence interval).

Format	Mean	Lower Bound	Upper Bound
A only	27.78	15.30	40.26
V only	32.50	20.18	42.16
AV	54.44	42.16	66.72
R_A (Auditory Enhancement)	.42	.32	.58

Table 2

Correlations of R_A with spoken word recognition and speech production tasks.

Test	Correlation with R_A	P value
MLNT Hard words	0.68	P<.05
LNT Easy words	0.78	P<.01
LNT Hard words	0.28	NS
PBK	0.28	NS
Speech Intelligibility	0.42	P<.05

Table 3

Mean percent correct scores for the two groups.

Group and Presentation Format	Single Talker, Easy Words	Single Talker, Hard Words	Multiple Talkers, Easy Words	Mulitple Talkers, Hard Words
CI: V	23.9	8.9	21.7	9.4
CI: A	34.4	29.4	38.6	23.9
CI: AV	75.8	64.2	70.0	52.7
NH: V	18.0	4.8	15.6	8.2
NH: A	54.2	45.2	48.9	39.7
NH: AV	75.2	70.9	74.3	62.2

Table 4

Mean visual enhancement (R_A) scores the two groups.

Group	Single Talker, Easy Words	Single Talker, Hard Words	Multiple Talkers, Easy Words	Mulitple Talkers, Hard Words
CI	.64	.49	.50	.35
NH	.40	.46	.49	.37